Effect of Slow Switching in On-line Learning for Ensemble Teachers

SEIJI MIYOSHI[1] * and MASATO OKADA[23]

[1]*Department of Electrical and Electronic Engineering, Faculty of Engineering Science, Kansai University, 3-3-35 Yamate-cho, Suita-shi Osaka, 564-8680*

[2]*Division of Transdisciplinary Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa-shi, Chiba, 277–8561*

[3]*RIKEN Brain Science Institute, 2–1 Hirosawa, Wako-shi, Saitama, 351–0198*

We have analyzed the generalization performance of a student which slowly switches ensemble teachers. By calculating the generalization error analytically using statistical mechanics in the framework of on-line learning, we show that the dynamical behaviors of generalization error have the periodicity that is synchronized with the switching period and the behaviors differ with the number of ensemble teachers. Furthermore, we show that the smaller the switching period is, the larger the difference is.

KEYWORDS: ensemble teachers, on-line learning, generalization error, statistical mechanics, slow switching

Learning can be classified into batch learning and on-line learning.[1,2] In on-line learning, examples once used are discarded and a student cannot give correct answers for all examples used in training. However, there are merits; for example, a large memory for storing many examples is not necessary and it is possible to follow a time variant teacher.[3,4] Recently, we used a statistical mechanical method[1,5] to analyze the generalization performance of a model composed of linear perceptrons: a true teacher, ensemble teachers, and the student in the framework of on-line learning.[6] That is, we treated a model that has K teachers called ensemble teachers who exist around a true teacher.[7] In the study, we analyzed the model in which a student switches the ensemble teachers in turn or randomly at each time step. Therefore, the study was an analysis of a fast switching model. On the contrary, the properties of a model in which a student switches the ensemble teachers slowly is also attractive. In this letter, we analyze such a slow switching model.

We have considered a true teacher, K ensemble teachers, and a student. They are all linear perceptrons with connection weights $\boldsymbol{A}$, $\boldsymbol{B}_k$, and $\boldsymbol{J}$, respectively. Here, $k = 1, \ldots, K$. For simplicity, the connection weight of the true teacher, the ensemble teachers, and the student is simply called the true teacher, the ensemble teachers, and the student, respectively. The true teacher $\boldsymbol{A} = (A_1, \ldots, A_N)$, ensemble teachers $\boldsymbol{B}_k = (B_{k1}, \ldots, B_{kN})$, student $\boldsymbol{J} = (J_1, \ldots, J_N)$, and input $\boldsymbol{x} = (x_1, \ldots, x_N)$ are N-dimensional vectors. Each component A_i of $\boldsymbol{A}$ is drawn

*E-mail address: miyoshi@ipcku.kansai-u.ac.jp

from $\mathcal{N}(0,1)$ independently and fixed, where $\mathcal{N}(0,1)$ denotes the Gaussian distribution with a mean of zero and a variance of unity. Some components B_{ki} are equal to A_i multiplied by –1, the others are equal to A_i. Which component B_{ki} is equal to $-A_i$ is independent from the value of A_i. Hence, B_{ki} also obeys $\mathcal{N}(0,1)$ and it is also fixed. The direction cosine between $\boldsymbol{B}_k$ and $\boldsymbol{A}$ is R_{Bk} and that between $\boldsymbol{B}_k$ and $\boldsymbol{B}_{k'}$ is $q_{kk'}$. Each of the components J_i^0 of the initial value $\boldsymbol{J}^0$ of $\boldsymbol{J}$ is drawn from $\mathcal{N}(0,1)$ independently. The direction cosine between $\boldsymbol{J}$ and $\boldsymbol{A}$ is R_J and that between $\boldsymbol{J}$ and $\boldsymbol{B}_k$ is R_{BkJ}. Each component x_i of $\boldsymbol{x}$ is drawn from $\mathcal{N}(0,1/N)$ independently.

This letter assumes the thermodynamic limit $N \to \infty$. Therefore, $\|\boldsymbol{A}\| = \|\boldsymbol{B}_k\| = \|\boldsymbol{J}^0\| = \sqrt{N}$, and $\|\boldsymbol{x}\| = 1$. Generally, norm $\|\boldsymbol{J}\|$ of the student changes as time step proceeds. Therefore, ratio l^m of the norm to $\sqrt{N}$ is introduced and called the length of the student. That is, $\|\boldsymbol{J}^m\| = l^m\sqrt{N}$, where m denotes the time step. The outputs of the true teacher, the ensemble teachers, and the student are $y^m + n_A^m$, $v_k^m + n_{Bk}^m$ and $u^m l^m + n_J^m$, respectively. Here, $y^m = \boldsymbol{A}\cdot\boldsymbol{x}^m$, $v_k^m = \boldsymbol{B}_k\cdot\boldsymbol{x}^m$, and $u^m l^m = \boldsymbol{J}^m\cdot\boldsymbol{x}^m$ where y^m, v_k^m, and u^m obey Gaussian distributions with a mean of zero and a variance of unity. n_A^m, n_{Bk}^m, and n_J^m are independent Gaussian noises with variances of $\sigma_A^2, \sigma_{Bk}^2$, and σ_J^2, respectively.

We define the error ϵ_{Bk} between true teacher $\boldsymbol{A}$ and each member $\boldsymbol{B}_k$ of the ensemble teachers by the squared errors of their outputs: $\epsilon_{Bk}^m \equiv \frac{1}{2}\left(y^m + n_A^m - v_k^m - n_{Bk}^m\right)^2$. In the same manner, we define error ϵ_{BkJ} between each member $\boldsymbol{B}_k$ of the ensemble teachers and student $\boldsymbol{J}$ by the squared errors of their outputs: $\epsilon_{BkJ}^m \equiv \frac{1}{2}\left(v_k^m + n_{Bk}^m - u^m l^m - n_J^m\right)^2$. Student $\boldsymbol{J}$ adopts the gradient method as a learning rule and uses input $\boldsymbol{x}$ and an output of one of the K ensemble teachers $\boldsymbol{B}_k$. Here, the student $\boldsymbol{J}$ uses each ensemble teacher $\boldsymbol{B}_k$ TN times succsessively where T is $O(1)$. That is,

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m - \eta\frac{\partial\epsilon_{BkJ}^m}{\partial\boldsymbol{J}^m} \tag{1}$$

$$= \boldsymbol{J}^m + \eta\left(v_k^m + n_{Bk}^m - u^m l^m - n_J^m\right)\boldsymbol{x}^m, \tag{2}$$

$$k = \mathrm{mod}\left(\left[\frac{m}{TN}\right], K\right) + 1, \tag{3}$$

where η denotes the learning rate and is a constant number. The Gauss notation is denoted by $[\cdot]$. That is, $\left[\frac{m}{TN}\right]$ is the maximum integer which is not larger than $\frac{m}{TN}$. Here, $\mathrm{mod}\left(\left[\frac{m}{TN}\right], K\right)$ denotes the remainder of $\left[\frac{m}{TN}\right]$ divided by K. Equation (3) means that the student uses each ensemble teacher $TN \sim O(N)$ times succsessively. We call this slow switching. By generalizing the learning rules, Eq. (2) can be expressed as $\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + f_k\boldsymbol{x}^m$, where f denotes a function that represents the update amount and is determined by the learning rule. In addition, we define the error ϵ_J between true teacher $\boldsymbol{A}$ and student $\boldsymbol{J}$ by the squared error of their outputs: $\epsilon_J^m \equiv \frac{1}{2}\left(y^m + n_A^m - u^m l^m - n_J^m\right)^2$.

One of the goals of statistical learning theory is to theoretically obtain generalization errors. Since generalization error is the mean of errors for the true teacher over the distribution

of new input and noises, generalization error ϵ_{Bkg} of each member $\boldsymbol{B}_k$ of the ensemble teachers and ϵ_{Jg} of student $\boldsymbol{J}$ are calculated as follows. Superscripts m, which represent the time step, are omitted for simplicity unless stated otherwise.

$$\begin{aligned}
\epsilon_{Bkg} &= \int d\boldsymbol{x} dn_A dn_{Bk} P\left(\boldsymbol{x}, n_A, n_{Bk}\right) \epsilon_{Bk} && (4)\\
&= \int dy dv_k dn_A dn_{Bk} P\left(y, v_k, n_A, n_{Bk}\right) \frac{1}{2}\left(y + n_A - v_k - n_{Bk}\right)^2 && (5)\\
&= \frac{1}{2}\left(-2R_{Bk} + 2 + \sigma_A^2 + \sigma_{Bk}^2\right), && (6)\\
\epsilon_{Jg} &= \int d\boldsymbol{x} dn_A dn_J P\left(\boldsymbol{x}, n_A, n_J\right) \epsilon_J && (7)\\
&= \int dy du dn_A dn_J P\left(y, u, n_A, n_J\right) \frac{1}{2}\left(y + n_A - ul - n_J\right)^2 && (8)\\
&= \frac{1}{2}\left(-2lR_J + l^2 + 1 + \sigma_A^2 + \sigma_J^2\right). && (9)
\end{aligned}$$

To simplify the analysis, two auxiliary order parameters $r_J \equiv lR_J$ and $r_{BkJ} \equiv lR_{BkJ}$ are introduced. Simultaneous differential equations in deterministic forms,[5] which describe the dynamical behaviors of order parameters when the student uses a teacher $\boldsymbol{B}_{k'}$ that consists of ensemble teachers have been obtained on the basis of self-averaging in the thermodynamic limits as follows:

$$\frac{dr_{BkJ}}{dt} = \langle f_{k'} v_k \rangle, \qquad \frac{dr_J}{dt} = \langle f_{k'} y \rangle, \qquad \frac{dl}{dt} = \langle f_{k'} u \rangle + \frac{1}{2l}\langle f_{k'}^2 \rangle. \tag{10}$$

Here, dimension N has been treated to be sufficiently greater than the number K of ensemble teachers. Time is defined by $t = m/N$, that is, time step m normalized by dimension N. Since linear perceptrons are treated in this letter, the sample averages that appeared in the above equations can be easily calculated as follows:

$$\langle f_{k'} u \rangle = \eta\left(\frac{r_{Bk'J}}{l} - l\right), \qquad \langle f_{k'}^2 \rangle = \eta^2\left(l^2 - 2r_{Bk'J} + 1 + \sigma_{Bk'}^2 + \sigma_J^2\right), \tag{11}$$

$$\langle f_{k'} y \rangle = \eta\left(R_{Bk'} - r_J\right), \qquad \langle f_{k'} v_k \rangle = \eta\left(q_{k'k} - r_{BkJ}\right). \tag{12}$$

Let us denote the values of r_J, r_{BkJ}, and l^2 of $t = t_0$ as $r_J^{t_0}, r_{BkJ}^{t_0}$, and $(l^2)^{t_0}$, respectively. By using these as intitial values, simultaneous differential equations Eqs.(10)–(12) can be solved analytically as follows:

$$\begin{aligned}
r_{BkJ} &= q_{k'k} + \left(r_{BkJ}^{t_0} - q_{k'k}\right) e^{-\eta(t-t_0)}, && (13)\\
r_J &= R_{Bk'} + \left(r_J^{t_0} - R_{Bk'}\right) e^{-\eta(t-t_0)}, && (14)\\
l^2 &= 1 + \frac{\eta}{2-\eta}\left(\sigma_{Bk'}^2 + \sigma_J^2\right) + 2\left(r_{Bk'J}^{t_0} - 1\right) e^{-\eta(t-t_0)}\\
&\quad + \left((l^2)^{t_0} - 1 - \frac{\eta}{2-\eta}\left(\sigma_{Bk'}^2 + \sigma_J^2\right) - 2\left(r_{Bk'J}^{t_0} - 1\right)\right) e^{\eta(\eta-2)(t-t_0)}. && (15)
\end{aligned}$$

Since all components A_i and J_i^0 of true teacher $\boldsymbol{A}$ and the initial student $\boldsymbol{J}^0$ are drawn

from $\mathcal{N}(0,1)$ independently, and because the thermodynamic limit $N \to \infty$ is also assumed, they are orthogonal to each other at $t = 0$. That is, $R_J = 0$ and $l = 1$ when $t = 0$.

In the following, we consider the case where direction cosines R_{Bk} between the ensemble teachers and the true teacher, direction cosines $q_{kk'}$ among the ensemble teachers and variances σ_{Bk}^2 of the noises of ensemble teachers are uniform. That is,

$$R_{Bk} = R_B, (k = 1, \ldots, K), \quad q_{kk'} = \begin{cases} 1, & (k = k'), \\ q, & (\text{otherwise}), \end{cases} \quad \sigma_{Bk}^2 = \sigma_B^2. \tag{16}$$

The dynamical behaviors of generalization errors ϵ_{Jg} have been analytically obtained by substituting Eqs. (14) and (15) into Eq. (9). The analytical results and the corresponding simulation results, where $N = 10^5$ are shown in Figs. 1 and 2. In computer simulations, ϵ_{Jg} was obtained by averaging the squared errors for 5×10^4 random inputs at each time step. In these figures, the curves represent theoretical results. The symbols represent simulation results. In these figures, $R_B = 0.7$ and $q = 0.49$ are common conditions. In addition, $\eta = 0.3, \sigma_A^2 = 0.1, \sigma_B^2 = 0.2$, and $\sigma_J^2 = 0.3$ are conditions for Fig. 1. $\eta = 1.5, \sigma_A^2 = 0.01, \sigma_B^2 = 0.02$, and $\sigma_J^2 = 0.03$ are conditions for Fig. 2.

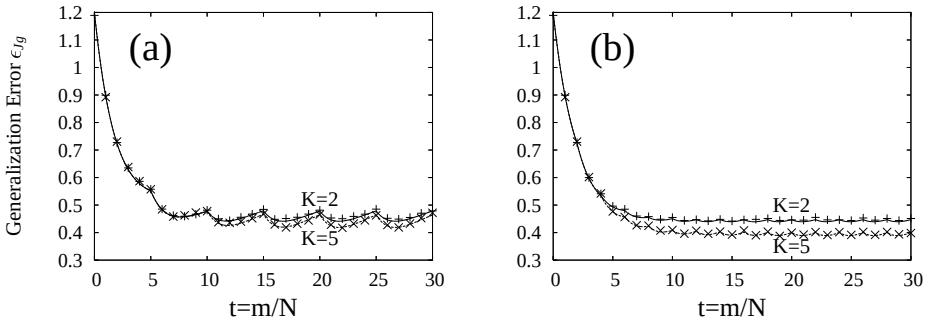

Fig. 1. Dynamical behaviors of generalization errors ϵ_{Jg} when $\eta = 0.3$. Theory and computer simulations. $R_B = 0.7, q = 0.49, \sigma_A^2 = 0.1, \sigma_B^2 = 0.2$, and $\sigma_J^2 = 0.3$. (a)$T = 5.0$, (b)$T = 2.0$.

These figures show that the dynamical behaviors of generalization error have the periodicity that is synchronized with the switching period T. In the case of $K = 2$, the student uses ensemble teachers as $\boldsymbol{B}_1 \to \boldsymbol{B}_2 \to \boldsymbol{B}_1 \to \boldsymbol{B}_2 \to \cdots$. In the case of $K = 5$, $\boldsymbol{B}_1 \to \boldsymbol{B}_2 \to \boldsymbol{B}_3 \to \boldsymbol{B}_4 \to \boldsymbol{B}_5 \to \boldsymbol{B}_1 \to \boldsymbol{B}_2 \to \boldsymbol{B}_3 \to \cdots$. Therefore, by comparing the behaviors of $K = 2$ and that of $K = 5$, the generarization errors ϵ_{Jg} completely agree during the time corresponding to two cycles from the initial state because the teachers used by student are the same. On the contrary, the generarization errors ϵ_{Jg} of $K = 2$ and $K = 5$ are not the same after the second cycle. In our study on the fast switching model,[6] it was proven that when a student's learning rate satisfies $\eta < 1$, the larger the number K is, the smaller the student's generalization error is. The same phenomenon is observed in the slow switching model treatd in this letter, that is, the generalization error of $K = 5$ is smaller than that of

$K = 2$ as shown in Fig. 1. On the contrary, the generalization error of $K = 5$ is larger than that of $K = 2$ in Fig. 2. Here, the dynamical behavior approaches that of the fast switching model[6] asymptotically in the limit of switching period $T \to 0$.

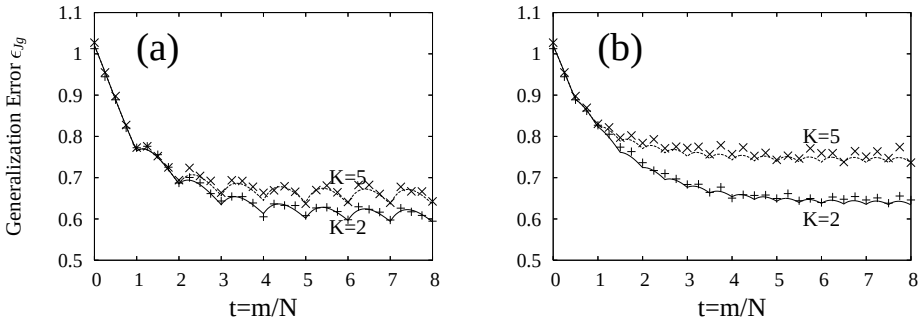

Fig. 2. Dynamical behaviors of generalization errors ϵ_{Jg} when $\eta = 1.5$. Theory and computer simulations. $R_B = 0.7, q = 0.49, \sigma_A^2 = 0.01, \sigma_B^2 = 0.02$, and $\sigma_J^2 = 0.03$. (a)$T = 1.0$, (b)$T = 0.5$.

In both cases of $\eta = 0.3$ and 1.5, the smaller the switching period T is, the larger the difference between the generalization error ϵ_{Jg} of $K = 2$ and that of $K = 5$ is. The reason is the following: if the switching period T is large, a student learns enough from only the one teacher that the student uses in the period. In other words, as the student forgets the other teachers, the influence of the number K of ensemble teachers becomes small.

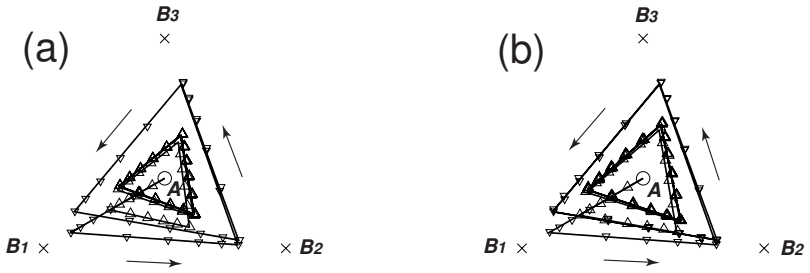

Fig. 3. Student's projection to 2-D plane on which $\boldsymbol{B}_1$–$\boldsymbol{B}_3$ exist. (a)$\eta = 0.3$, (b)$\eta = 1.5$. Solid lines represent trajectories of student's projection obtained theoretically. Symbols $\triangle$ and $\bigtriangledown$ represent computer simulations with (a)$T = 2.0$ and $T = 5.0$, (b)$T = 0.5$ and $T = 1.0$, respectively.

We visualize the student's behaviors in the case of $K = 3$ to understand them intuitively. That means we obtain the student's projection to the two-dimensional plane on which the three ensemble teachers exist. Figure 3 shows the projection's trajectories in the case of $\eta = 0.3$ and $\eta = 1.5$. In this figure, symbols $\times$, $\circ$ and solid lines represent the ensemble teachers $\boldsymbol{B}_1$, $\boldsymbol{B}_2$ and $\boldsymbol{B}_3$, the projection of the true teacher $\boldsymbol{A}$ and the trajectories of the student's projection

obtained theoretically, respectively. In Fig. 3(a), symbols $\triangle$ and $\bigtriangledown$ represent the student's projections obtained by computer simulations with $T = 2.0$ and $T = 5.0$, respectively. In Fig. 3(b), those represent the projections with $T = 0.5$ and $T = 1.0$, respectively. This figure shows that the student moves straight toward the teacher that the student uses then. Therefore, the student's trajectories in the steady state are regular triangles. The triangles are small when the switching period T is small and the triangles are large when T is large. In this figure, a side of the trajectory corresponds to a period in Figs. 1 and 2. Note that the distance between the student and the true teacher in Fig. 3 is not necessarily related to the real distance between the student and the true teacher nor the generalization error since this figure shows the projections. Though the student is near the true teacher when T is small in Fig. 3(b), the generalization error is small when T is large as shown in Fig. 2.

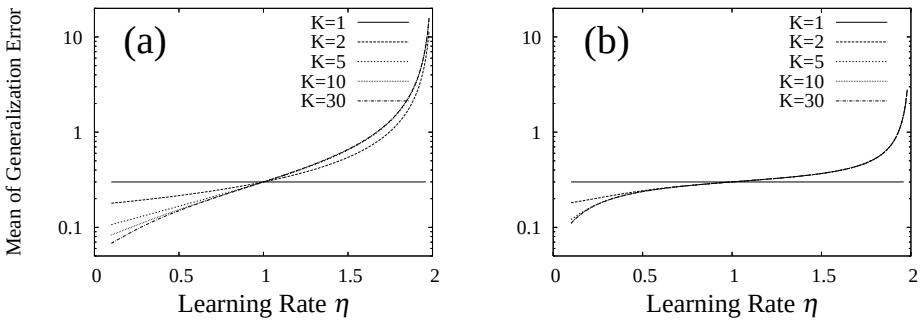

Fig. 4. Means of steady state generalization errors ϵ_{Jg}. Theory. $q = 0.49, R_B = 0.7$ and $\sigma_A^2 = \sigma_B^2 = \sigma_J^2 = 0.0$. (a)$T = 0.5$, (b)$T = 5.0$.

The relationships between the learning rate η and the means of steady state generalization errors ϵ_{Jg} are shown in Fig. 4. The means are measured by averaging the generalization errors during a cycle after the dynamical behaviors reach the steady state. In this figure, when a learning rate satisfies $\eta < 1$, the larger the number K is, the smaller the generalization error is. This is the same property with that of the fast switching model.[6] A comparison of Figs. 4(a) and 4(b) shows that the smaller the switching period T is, the larger the difference among the means of generalization errors ϵ_{Jg} of various K values in the slow switching model as treated in this letter.

The relationships between the learning rate η and the means of steady state generalization errors ϵ_{Jg} for various direction cosines q are shown in Fig. 5. As shown in this figure, when a learning rate satisfies $\eta < 1$, the smaller q is, the smaller the generalization error is. This is also the same property as that of the fast switching model.[6] By comparing Figs. 5(a) and 5(b), we see that the smaller the switching period T is, the larger the difference among the means of generalization errors ϵ_{Jg} of various q.

In summary, we have analyzed the generalization performance of a student in a model

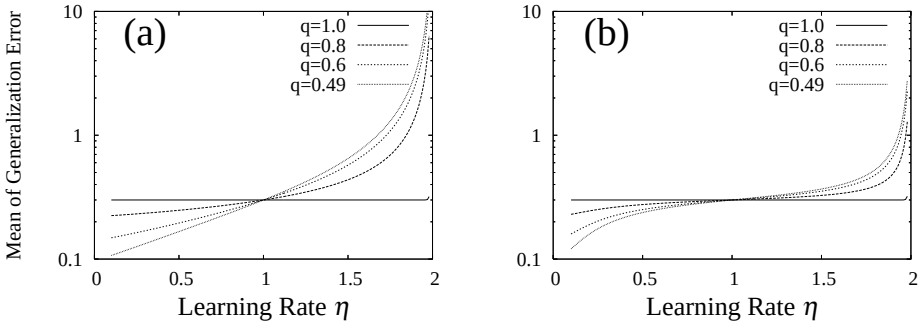

Fig. 5. Means of steady state generalization errors ϵ_{Jg}. Theory. $K = 5, R_B = 0.7$ and $\sigma_A^2 = \sigma_B^2 = \sigma_J^2 = 0.0$. (a)$T = 0.5$, (b)$T = 5.0$.

composed of linear perceptrons: a true teacher, ensemble teachers, and the student. In particular, the case where the student slowly switches ensemble teachers has been analyzed. By calculating the generalization error analytically using statistical mechanics in the framework of on-line learning, we have shown that the dynamical behaviors of generalization error have the periodicity that is synchronized with the switching period and that the behaviors differ with the number of ensemble teachers. Furthermore, we have shown that the smaller the switching period is, the larger the difference is.

Acknowledgments

This research was partially supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan, with Grants-in-Aid for Scientific Research 16500093, 18020007, 18079003, and 18500183.

References

1) D. Saad, (ed.): *On-line Learning in Neural Networks* (Cambridge University Press, Cambridge, 1998).
2) N. Cesa-Bianchi and G. Lugosi: *Prediction, Learning, and Games* (Cambridge University Press, New York, 2006).
3) S. Miyoshi and M. Okada: J. Phys. Soc. Jpn. **75** (2005) 024003.
4) M. Urakami, S. Miyoshi, and M. Okada: J. Phys. Soc. Jpn. **76** (2005) 044003.
5) H. Nishimori: *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press, Oxford, 2001).
6) S. Miyoshi and M. Okada: J. Phys. Soc. Jpn. **75** (2006) 044002.
7) T. Hirama and K. Hukushima: J. Phys. Soc. Jpn. **77** (2008) 094801.